INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

State-of-the-art MRI Protocol for Comprehensive Assessment of Vocal Tract Structure and Function

Sajan Goud Lingala[1], *Asterios Toutios*[1], *Johannes Toger*[1], *Yongwan Lim*[1], *Yinghua Zhu*[1], *Yoon-Chul Kim*[2], *Colin Vaz*[1], *Shrikanth Narayanan*[1], *Krishna Nayak*[1]

[1]Electrical Engineering, University of Southern California, Los Angeles, CA, USA
[2]Samsung Medical Center, Seoul, South Korea

lingala@usc.edu, knayak@usc.edu

Abstract

Magnetic Resonance Imaging (MRI) provides a safe and flexible means to study the vocal tract, and is increasingly used in speech production research. This work details a state-of-the-art MRI protocol for comprehensive assessment of vocal tract structure and function, and presents results from representative speakers. The system incorporates (a) custom upper airway coils that are maximally sensitive to vocal tract tissues, (b) graphical user interface for 2D real-time MRI that provides on-the-fly reconstruction for interactive localization, and correction of imaging artifacts, (c) off-line constrained reconstruction for generating high spatio-temporal resolution dynamic images at (83 frames per sec, 2.4 mm^2), (d) 3D static imaging of sounds sustained for 7 sec with full vocal tract coverage and isotropic resolution (resolution: 1.25 mm^3), (e) T2-weighted high-resolution, high-contrast depiction of soft-tissue boundaries of the full vocal tract (axial, coronal, sagittal sweeps with resolution: 0.58 x 0.58 x 3 mm^3), and (f) simultaneous audio recording with off-line noise cancellation and temporal alignment of audio with 2D real-time MRI. A stimuli set was designed to capture efficiently salient, static and dynamic, articulatory and morphological aspects of speech production in 90-minute data acquisition sessions.

Index Terms: MRI system for speech production, constrained reconstruction, rapid real-time MRI, structural and functional characterization of vocal tract

1. Introduction

Speech production involves a complex coordination of several upper and lower airway organs. Magnetic Resonance Imaging (MRI) is increasingly used in speech production research. In contrast to modalities such as electro-magnetic articulography (EMA), X-rays, and ultrasound, MRI provides a safe and flexible means to study the vocal tract with excellent soft-tissue contrast, capability to view deep structures, and flexibility to view in any imaging plane [1, 2, 3]. MRI however has been challenged by slow imaging speed, which places a challenging trade-off amongst the achievable spatio-temporal resolutions, slice coverage, and signal to noise. For instance, early use of MRI in speech production research utilized dynamic imaging in "cine" mode, which utilized several repetitions of a speech task to synthesize retrospectively a dynamic cine loop [2]. Advances in rapid non-Cartesian imaging have enabled real-time depiction of articulatory dynamics up to a native temporal resolution of 78 ms/frame, and has enabled several insights in speech science and language production [1, 4]. In this work, we put together improved upper-airway MRI acquisition and reconstruction imaging tools that improves MRI imaging trade-offs [5, 6], and propose a state-of-the art 90-minute protocol for comprehensive assessment of the vocal tract structure and function. The proposed system synergistically combines advances in custom coil design, sparse-sampling, constrained reconstruction, and simultaneous audio acquisition to enable a) rapid 2D real time MRI at 2.4 mm^2/pixel, and 12 ms/frame, b) 3D imaging of the full-vocal tract with isotropic 1.25 mm^3 resolution in 7 sec, and also utilizes c) conventional T2-weighted sequences to provide high soft-tissue contrast images. The stimuli in the proposed protocol was designed to capture efficiently salient, static and dynamic, articulatory and morphological aspects of speech production.

2. MRI protocol for comprehensive vocal tract imaging

2.1. Upper-airway custom coils

All our experiments are performed on a 1.5 T GE Signa Excite scanner with high performance gradients (40 mT/m amplitude, and 150 mT/m/ms slew rate). We utilize custom designed coils for upper-airway imaging. The coil geometry has eight coil elements with four on either side of the jaw. The rationale for choosing custom coils is the superior sensitivity to all upper airway regions of interests including tongue, lips, and also deep structures such as velum, epiglottis, and glottis. We have shown that the custom coil provides a signal to noise ratio boost of 2-6 fold in various upper-airway regions in comparison to coil arrays developed for other body parts (eg. head coils, head and neck coils) [5]. The coil geometry is designed in a way that there is an opening near the mouth so that the microphone can be positioned in close proximity to the mouth (see Fig 1). The coils are available in two sizes for adults and children respectively.

2.2. Rapid 2D real-time MRI to evaluate dynamics of vocal tract

Real-time 2D imaging was performed via a custom real-time interactive imaging platform (RT-Hawk, Heart Vista Inc, Los Altos, CA) [7]. A multi-shot short spiral readout spoiled gradient echo pulse sequence (flip angle: 15^0, slice thickness: 6 mm; readout time: 2.5msec, repetition time (TR): 6.004 msec, spatial resolution: 2.4 mm^2) was implemented.

http://dx.doi.org/10.21437/Interspeech.2016-559

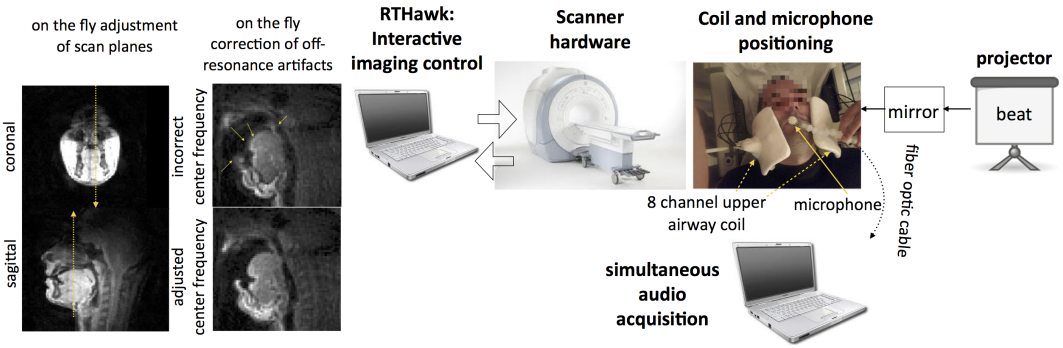

Figure 1: *Subject positioning and imaging set up: The custom 8-channel upper-airway coil is positioned in close proximity to the upper-airway structures, and has an opening near the mouth for microphone positioning. The speaker visualizes the stimuli through a mirror projector set up. The real-time Hawk interface (Heart Vista Inc, Los Altos, CA, USA) provides an interactive interface with the scanner hardware and has flexibility for real-time adjustment of scan planes, and correction of image artifacts. Audio is recorded simultaneously, and is processed off-line to provide MRI acoustic noise-cancelled recordings.*

2.2.1. *On-the-fly image reconstruction*

On-the fly image reconstruction was implemented within RT-Hawk using a fast implementation of the gridding algorithm (i.e, interpolation of spiral samples onto a Cartesian grid, followed by inverse Fourier transform). The images were formed on-the-fly with a time resolution of 78 msec. The minimal latency of the reconstruction ($<$100 msec) allowed for instant feedback to the operator and enabled efficient adjustment of scan planes during localization, and on-the-fly correction of imaging artifacts such as the off-resonance blurs at air-tissue interfaces. Specifically, the subject was asked to open their mouth, and the operator qualitatively adjusted the center frequency such that the air-tissue interfaces are sharp (majorly at the air-tongue, air-velum, air-lip).

2.2.2. *Off-line constrained reconstruction*

To enable improved time resolution in the dynamic images, an off-line sparse-SENSE constrained reconstruction algorithm was implemented [5]. A sparsity based temporal finite difference constraint which assumes the desired information is contained in the moving edges was considered. Specifically, image reconstruction was formulated as the following convex minimization problem:

$$\min_{f(r,t)} \|\mathcal{A}(f) - \mathbf{b}\|_2^2 + \lambda \|\nabla_t(f)\|_1 \qquad (1)$$

where $\mathbf{b}$ is a concatenated vector containing the spiral noisy k-t measurements. $\mathcal{A}$ is the forward model that models for coil sensitivity encoding, and Fourier transform on a specified spiral trajectory in each time frame; $f(r,t)$ denotes the dynamic image time series; $r\epsilon(x,y)$. ∇_t denotes a temporal finite difference operator, and $\|.\|_1$ denotes the l_1 norm that is defined as sum of absolute values of the entries. Two spiral interleaves were considered in each time frame, which resulted in the final reconstructed dynamic images to have a time resolution of 12 ms/frame (83.33 frames per sec). The regularization parameter λ balances the trade-off between the constraint and data-fidelity terms, and was chosen empirically as $\lambda = 0.002$. The optimization problem in (1) was solved using a recent open-source Berkeley Advanced Reconstruction Tool Box (BART) [8]; the reconstruction times were on the order of 34 minutes to reconstruct a 30-second speech snippet ($n_x \times n_y \times t = 84 \times 84 \times 2500$) using 8 cores on a 16-core Intel(R) Xeon(R) CPU E5-2698 v3; 2.30GHz, with 40 MB of L3 cache.

Table 1: *Comprehensive vocal-tract imaging protocol*

Purpose	Index	Task	Length
Real-time 2D MRI (scripted speech)	R1-R3	Consonants in symmetric VCV (3 scans)	30 sec (x3)
	R4	Vowels in bVt	30 sec
	R5	four phonetically rich sentences [9]	30 sec
	R6	Rainbow passage [10]	30 sec
	R7	Grandfather passage [11]	30 sec
	R8	Northwind and the sun passage [12]	30 sec
	R9	Gestures	30 sec
	R10-R18	Repetition of index R1-R9	30 sec (x9)
Real-time 2D MRI (spontaneous speech)	S1-S5	Description of pictures	30 sec (x5)
	S6-S10	Questions/ Discussion topics	30 sec (x5)
Accelerated 3D volumetric static MRI of the full vocal tract	V1	Vowels, continuant consonants, postures	7 sec (x33)
	V2	Repetition of V1	7 sec (x33)
T2-weighted scans for anatomical reference	Sagittal	Resting posture	$\sim$3.5 min
	Coronal	Resting posture	$\sim$3.5 min
	Axial	Resting posture	$\sim$3.5 min

2.3. Static high-resolution imaging of the full vocal tract at 7 seconds during sustained sounds and postures

To characterize the vocal tract shape during sustained sounds, and various postures, an accelerated 3D gradient echo sequence with sparse-sampling was implemented to image the full vocal tract with isotropic resolution [6]. Sequence parameters were: TR=3.8 msec; flip angle=5^0; spatial resolution: $1.25\times1.25\times1.25$ mm^3; field of view: $20\times20\times10$ cm^3 respectively in the anterior-posterior (A-P), superior-inferior (S-I), and left-right (L-R) directions; image matrix size: $160\times160\times80$. The center portion (40×20) of the ky-kz space was fully sampled to estimate the coil sensitivities from the data itself. The outer portions of the ky-kz space was sampled using a sparse Poisson-Disc sampling pattern, which together resulted in a net acceleration of 7 fold, and a total scan time of 7 seconds. Image

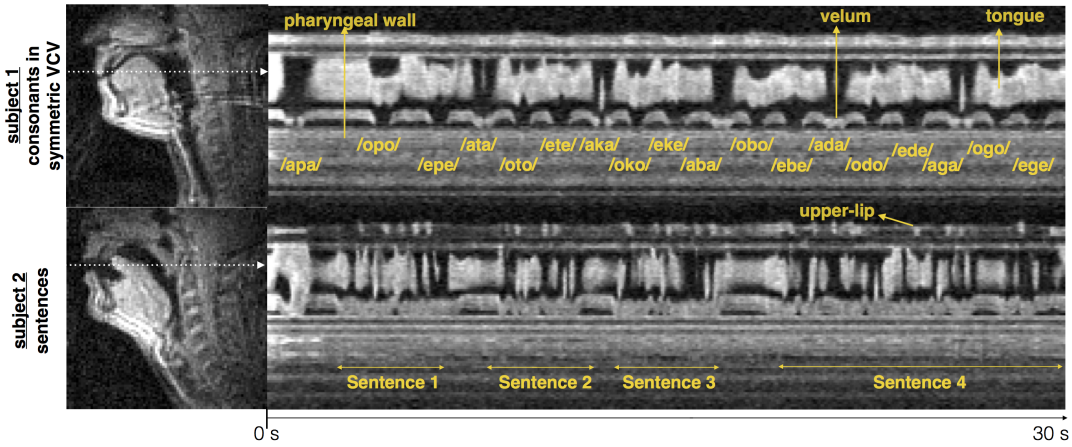

Figure 2: *Demonstration of the speech stimuli of (a) producing consonants in symmetric VCV phrases, and (b) producing scripted sentences on two different subjects. The horizontal dotted arrows in the first column correspond to image cuts along which the temporal profiles are shown in the second column. The proposed system depicts the articulatory dynamics of the various speech stimuli with good temporal fidelity. The improved temporal resolution (83 frames/sec) shows flexibility in adapting to capture fast arbitrary articulatory motion in different speech tasks, and also adapts to subject specific speech rates (eg. compare speech rate in second row v/s first row). The improved spatial resolution (2.4mm^2/pixel) ensures improved capture and depiction of dynamics of small structures such as the epiglottis.*

reconstruction was achieved off-line by a sparse-SENSE constrained reconstruction convex minimization problem, which utilizes sparsity based spatial finite difference constraints:

$$\min_{f(r,t)} \|\mathcal{A}(f)-\mathbf{b}\|_2^2+\lambda \left\| \sqrt{|\nabla_x(f)|^2+|\nabla_y(f)|^2+|\nabla_z(f)|^2} \right\|_1 \quad (2)$$

where the first term denotes data-consistency, where $\mathcal{A}$ models for Fourier under-sampling and coil sensitivity encoding, and the sparsity based total variation regularization term penalizes rapidly changing pixel intensities (corresponding to under-sampling artifacts), while preserving high-resolution edge information at various air-tissue interfaces. λ balances the trade-off between the regularization and data-fidelity terms, and was chosen empirically as $\lambda = 0.2$. (2) was solved using the BART tool box [8]; the reconstruction time was ≈45 seconds on a 16-core Intel(R) Xeon(R) CPU E5-2698 v3; 2.30GHz, with 40 MB of L3 cache.

2.4. T2-weighted MRI for structural characterization of vocal tract

High-resolution, high-contrast T2-weighted fast spin-echo based sequence was considered to provide images with soft tissue contrast. The rationale for this sequence was to clearly identify soft-tissue boundaries and is included as a means to aid segmentation (manual or automatic) of the vocal organs. Sequence parameters were: TR: 4600ms, TE:120ms; slice thickness: 3 mm; in-plane resolution: 0.58x0.58 mm^2; in-plane field of view: 30x30 cm^2; number of averages: 1; echo train length: 25; scan time: 3.5 minutes. The sequence was run to obtain full sweeps of the vocal tract in the axial, sagittal, and coronal orientations.

2.5. Audio acquisition

Audio was recorded concurrently with MRI acquisition inside the MRI scanner while subjects are imaged, using a fiberoptic microphone (Optoacoustics Ltd., Moshav Mazor, Israel) and custom recording and synchronization setup [13]. Speech in the recorded audio was then enhanced, using a customized denoising method [13], in order to reduce the effect of loud scanner noise.

2.6. 90-minute protocol to study structural and functional aspects of the vocal tract

Table 1 shows a comprehensive 90-minute protocol with the proposed sequences, where the stimuli set was designed to capture efficiently salient, static and dynamic, articulatory and morphological aspects of speech production. Rapid real-time 2D MRI in the mid-sagittal view was used to obtain recordings of scripted speech and spontaneous speech, where the stimuli in scripted speech was repeated twice.

The scripted speech contained producing consonants in symmetric vowel-consonant-vowel context, vowels interspersed between consonant b and t in b-V-t context, four phonetically rich sentences [9], and three commonly used passages in linguistic studies [10, 11, 12]. Several gestures such as clenching, wide opening of mouth and yawning, swallowing, sustained sound production of the sounds "eee-aah-uuw-eee", tracing of palate with tongue tip, singing "la" at highest, lowest note, were also acquired and repeated with 2D real-time MRI. Spontaneous speech tasks involved describing contexts in five randomly shown pictures, and discussion of five general topics (eg. "what is your favorite restaurant"). All the 2D real time scans involved pause time of ≈30 seconds between stimuli to allow enough recovery for the subject prior to the next task, and also to avoid gradient heating.

With the accelerated volumetric protocol, we acquire two repetitions of 33 stimuli which contained producing sustained sounds of the vowels, continuant consonants, and several postures such as normal breathing with mouth closed, clenching of teeth, sticking of tongue out as far as possible, pulling back the tongue in as far as possible, tongue tip raise to the middle of the palate, and holding breath. In the accelerated volumetric protocol, a recovery time of ≈5-10 seconds was given to the subject between the stimuli. The final set of scans involve acquisition of

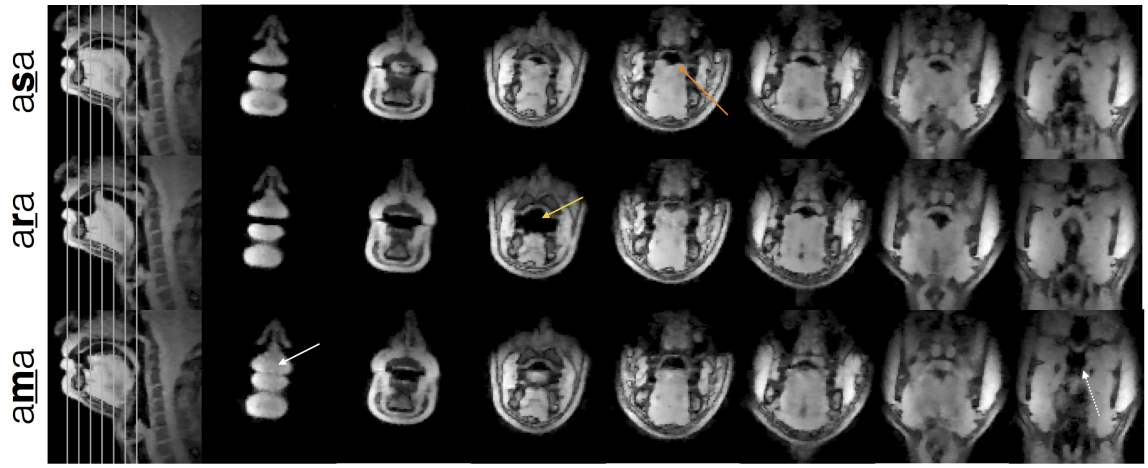

Figure 3: *3D high-resolution imaging of the full vocal tract during production of sounds sustained for 7 seconds: Shown here are examples of various sustained continuant consonants. The first column depicts the mid-sagittal view from the dataset, and columns two to eight depict coronal views from anterior to posterior as marked by the vertical lines in the first column.*

T2-weighted images in the resting posture. Including the pause times between the stimuli, the total acquisition time for the protocol was about 90 minutes.

3. Image quality examples

Figure 2 demonstrates example image quality from two representative speakers with the 2D rapid real-time MRI protocol (spatio-temporal resolutions of 2.4 mm^2 and 12 ms/frame). The images depict high spatio-temporal fidelity in depicting the various articulatory dynamics. The proposed constrained reconstruction framework exploits local redundancies amongst local temporal frames, and therefore is adaptive to a range of speech tasks such as those shown in Figure 2.

Figure 3 demonstrates example image quality with the 3D accelerated full vocal tract coverage, 1.25 mm^3 isotropic resolution protocol. As shown by the different views, the proposed protocol allows to distinguish subtle differences of the articulators amongst different sounds (see arrows in Fig.3). For instance, tongue grooving during /asa/ is more pronounced in the coronal view, the closure of lips during /ama/ can be depicted in two dimensions in the coronal view, the airway opening and closing near the tongue tip during /ara/ can be characterized through the coronal sweep from anterior to posterior.

Figure 4 demonstrates example image quality from the T2-weighted scans. The high resolution and high-contrast information in these images demonstrate clear depiction of various soft-tissue boundaries, which aids in segmentation of vocal tract organs.

4. Discussion

We have developed a state of the art MRI protocol for comprehensive analysis of various structural and functional aspects of the vocal tract. A synergetic combination of custom upper-airway coil design, sparse-sampling, and constrained reconstruction enables improved imaging for rapid real time 2D MRI at upto 83 frames/sec, and accelerated 3D high-resolution full vocal tract coverage in 7 seconds. High-resolution, high-soft tissue contrast T2-weighted images are also acquired. A comprehensive 90 minute protocol with an efficiently designed

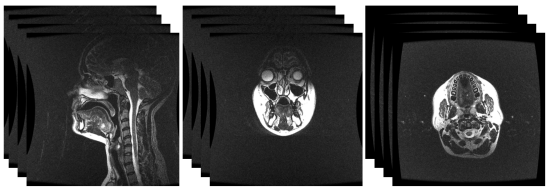

Figure 4: *T2-weighted images for high contrast, high resolution depiction of soft-tissue boundaries. Full sweeps of the sagittal, coronal, and axial views are shown . The utility of these images are to identify clear depiction of the soft-tissue boundaries and is a means to aid segmentation of the vocal organs.*

speech stimuli set is proposed to capture various functional and structural aspects of the vocal tract. The proposed protocol allows to gain insights into several aspects of speech production such as characterization of how different speakers who have different sizes and shapes of the vocal tract produce different speech tasks, identification of differences between speech tasks within a speaker, and many more.

Some aspects of the imaging methodology can be further improved. First, the reconstruction time of 2D real-time data sets (≈34 mins to reconstruct a 30 sec speech sample) has scope for further improvement by exploiting appropriate code parallelism, and using graphical processing units. Secondly, the regularization parameter in constrained reconstruction was chosen empirically in this study - however, a current focus of our investigation is to devise quantitative criterion in choosing the regularization parameters. Lastly, test/re-test studies of both the improved real-time 2D and accelerated 3D imaging protocols is a subject of our ongoing research, where we aim to quantify the repeatability and reproducibility of quantitative measures derived from the imaging data.

5. Acknowledgements

We acknowledge funding support from NIH-R01-DC007124 and NSF-1514544.

6. References

[1] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.

[2] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech mri: morphology and function," *Physica Medica*, vol. 30, no. 6, pp. 604–618, 2014.

[3] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech mri," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, 2016.

[4] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[5] S. G. Lingala, Y. Zhu, Y.-C. Kim, A. Toutios, S. Narayanan, and K. S. Nayak, "A fast and flexible mri system for the study of dynamic vocal tract shaping," *Magnetic resonance in medicine*, vol. early view, Jan 17, no. doi: 10.1002/mrm.26090, 2016.

[6] Y.-C. Kim, S. S. Narayanan, and K. S. Nayak, "Accelerated three-dimensional upper airway mri using compressed sensing," *Magnetic Resonance in Medicine*, vol. 61, no. 6, pp. 1434–1440, 2009.

[7] J. M. Santos, G. A. Wrigh, and J. M. Pauly, "Flexible real-time magnetic resonance imaging framework," in *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 1. IEEE, 2004, pp. 1048–1051.

[8] M. Uecker, F. Ong, J. I. Tamir, D. Bahri, P. Virtue, J. Y. Cheng, T. Zhang, and M. Lustig, "Berkeley advanced reconstruction toolbox," in *Proceedings of the 23rd Annual Meeting ISMRM, Toronto*, 2015, p. 2486.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[10] G. Fairbanks, "The rainbow passage," *Voice and articulation drillbook*, vol. 2, 1960.

[11] A. E. Aronson and J. R. Brown, *Motor speech disorders*. WB Saunders Company, 1975.

[12] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[13] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.